Noname manuscript No.
(will be inserted by the editor)

Evolutionary Linkage Creation between Information Sources in P2P Networks

Kei Ohnishi · Mario Köppen · Kaori Yoshida

Received: date / Accepted: date

Abstract The present paper proposes a peer-to-peer (P2P) information retrieval and sharing system that evolutionarily creates linkages of information sources that are useful for both information publishers and information users, where information is managed in a decentralized manner. The proposed system relies on interactions among information publishers who actually generate information and have the greatest knowledge of the information, information users who use the information, and a network that creates useful linkages of information sources (information publishers). In order to enhance the value of their own information sources, information publishers propose new linkages of information sources that indicate information sources with which they would like to have their own information sources co-occur. The information users evaluate the linkages proposed by the information publishers. The network evolutionarily reconstructs the topological structures of the P2P network based on the fitness obtained from the users. Simulation results suggest that it is possible to find more information sources that users desire using the topological structures reconstructed by the proposed system, as compared to the use of non-reconstructed topological structures.

Keywords Information sources · P2P networks · Evolutionary algorithms · Information retrieval

1 Introduction

The amount of information on the Internet is now rapidly increasing due to rapid growth of technology for easy publishing and sharing information, as well as information, communication, and data storage technology. In this information explosion era, how information search and retrieval systems provide useful information for information users (just called "users" hereinafter) is important.

In information search and retrieval systems on a client-server network such as the Web, the locations of information such as documents, images, music, movies, and so on in the systems are managed by servers. In this case, the systems need not be concerned with the locations of information. The primary focus of the present study is how to assign identifiers to information. One method is to have only authorized mechanisms to assign identifiers to information in an integrated manner. Another method is to have several general users freely assign identifiers to information. Such an information search and retrieval system is referred to as a folksonomy [1][2][3].

Meanwhile, recently, peer-to-peer (P2P) networks [4] have attracted great attention. Unlike client-server networks, P2P networks do not fix a role of node, and every node can be both server and client. In addition, since nodes that can be both client and server provide some service for each other in a P2P network, a P2P network can easily and quickly start some service among joining nodes aside from its service quality and scale, which is one of the strong points of P2P networks. However, in P2P networks, a search mechanism is a must, no matter what objects are searched. That is because P2P networks do not have a centralized mechanism for managing locations of what nodes search the networks for as client-server networks have. More accurately, P2P networks that allow node to freely make links to other nodes are referred to as unstructured P2P networks and in unstructured P2P networks, there is no mechanism to manage locations of objects for which nodes look. One of the representative unstructured P2P networks is the Gnutella [5][6].

In unstructured P2P networks, linkages between information sources (nodes) that are created by humans can pro-

K. Ohnishi, M. Köppen, and K. Yoshida
Kyushu Institute of Technology,
680-4 Kawazu, Iizuka-shi, Fukuoka 820-8502, JAPAN
Tel. & Fax: +81-948-29-7660
E-mail: {ohnishi@cse, mario@ndrc, kaori@ai}.kyutech.ac.jp

vide good clues for obtaining useful information sources. For instance, on the present-day Web, hyperlinks whose labels, for example “my recommended hospital”, can express the meanings of the links correspond to such linkages and enable users to navigate among useful and trusted web sites. In addition, information publishers (just called “publishers” hereinafter) in unstructured P2P networks, who have the greatest knowledge concerning the information that they have published, should be able to contribute to information search by proposing linkages among information sources, including their own, from their own unique perspective. Actually, information sources in unstructured P2P networks are allowed to freely make network links to others and the network links are regarded as a sort of linkages between information sources. Search queries issued in unstructured P2P networks are forwarded on the network links in some ways such as a random walk based method. However, in this case, the information sources cannot express the meanings of linkages, and more describable linkages are required to facilitate to obtain useful information sources .

In the present paper, we propose a P2P information retrieval and sharing system that creates linkages of information sources that are useful for both publishers and users, where information is managed in a decentralized manner. The proposed system relies on interactions among publishers who actually generate information and have the greatest knowledge of the information, users who use the information, and a network that creates useful linkages of information sources (publishers). In addition, through simulations, we evaluate the proposed system that includes a cycle of three procedures, which are (1) the proposal of linkages between information sources by publishers, (2) the evaluation of the proposed linkage of the information sources by users, and (3) the reconstruction of linkages between information sources by a network.

The remainder of the present paper is organized as follows. We briefly describe related research in Section 2. In Section 3, we describe the evolutionary approach to creating linkages between information sources for efficient P2P information retrieval and sharing. Section 4 shows the results of evaluation of the proposed approach. Finally, Section 5 presents conclusions and describes areas for future research.

2 Related Work

The proposed P2P information retrieval and sharing system evolutionarily modifies its several co-existing topological structures based on the fitness obtained from users in order to improve the efficiency of information retrieval and sharing. In addition, the proposed system is characterized by a mechanism to return not only information fitting a query issued by a user, but also additional information offered by publishers to the user. We describe the related work from these two perspectives.

We have proposed a P2P networking technique that dynamically and evolutionarily optimizes several co-existing topological structures of an unstructured P2P network based on the fitness obtained from the P2P nodes (users) [7]. This technique uses an evolutionary algorithm [8], which is a general term for a meta-heuristics optimization algorithm inspired by biological genetics and evolution, for optimizing topological structures of an unstructured P2P network. There has been no research on such an evolutionary P2P networking technique. However, several methods for dynamic modification of a single topological structure of a P2P network have been reported [9][10][11]. These methods basically reconstruct local topological structures using local information on the state of the network.

An evolutionary algorithm has been used to optimize the parameter values of a P2P network using fitnesses obtained from a simulation model of the P2P network [12]. Unlike EP2P used in this paper, this is not an online approach to optimizing the parameters of P2P networks. In addition to P2P networks, an evolutionary algorithm has been applied to on-line optimization of communication networks, such as on-line optimization of routing tables of routers in the Internet [13] and that of protocol stacks [14].

Information retrieval and recommendation systems that return additional information to a user who has issued a search query have been developed for the Web on a client-server network, such as a system implementing collaborative filtering [15]. In addition, hyperlinks on the Web are actually linkages between information sources that can be created by publishers. However, there is no such a mechanism in distributed systems, such as P2P networks.

3 Evolutionary Creation of Linkages between Information Sources

3.1 Concept

The P2P information retrieval and sharing system presented herein is based on the evolutionary P2P networking technique (EP2P) [7], which dynamically and evolutionarily optimizes several topological structures of a P2P network that includes all nodes, using the fitness obtained from the P2P nodes (users). In fact, we use a genetic algorithm [16] here, which is one type of evolutionary algorithm, for optimizing the topological structures. A genetic algorithm has the same general flow as an evolutionary algorithm but relies mainly on a crossover (recombination) operator to find better solutions. However, we keep using terms of “evolutionary algorithm” and “evolutionary operators” below. A general flow chart of EP2P is shown in Figure 1. The system

presented herein attempts to create linkages between information sources (P2P nodes) that are useful for both publishers and users based on interactions among publishers, users, and a P2P network executing EP2P (see Figure 2). In other words, the linkage creation between information sources is conducted as the third and the fourth procedures in the general flow chart of EP2P shown in Figure 1. We hereinafter refer to the proposed P2P information retrieval and sharing system as the linkage creating evolutionary P2P system for information retrieval and sharing (L-EP2P).

The linkage of information sources in L-EP2P basically represents information sources (P2P nodes) that are obtained from search queries that are issued by a user and several publishers, as well as directed links for query propagation in topological structures of a P2P network. Publishers are equivalent to information sources (P2P nodes) that are included in several coexisting topological structures of a P2P network, which are dynamically and evolutionarily modified by EP2P. Users are allowed to freely create a fixed number of links to publishers (P2P nodes) for issuing search queries, but these links are not modified by EP2P. Publishers are participants who hold their own information, and their roles are not only to publish their own information but also to propose linkage of information sources. Users are also participants who only search the P2P network for desired information.

As shown in Figure 2, L-EP2P consists of publishers, users, and EP2P. EP2P consists of a P2P network, including several co-existing topological structures and a super node that executes an evolutionary algorithm to modify the co-existing topological structures of the P2P network. It is also possible to use several super nodes that back each other up so that EP2P will be fault-tolerant and scalable.

The purpose of L-EP2P is to create linkages among information sources that are useful for both publishers and users. The mechanism by which to achieve these linkages is an enduring circulation of three procedures: (1) publishers propose linkages of information sources that represent other information sources that should co-occur with their own information sources in order to enhance the value of their own information sources, (2) users evaluate the linkages of information sources proposed by the publishers based on actual results of information search, and (3) EP2P reconstructs the linkages of information sources realized by directed P2P network links using the fitness obtained from the users. The circulation of the above three procedures corresponds to interactions among publishers, users, and a P2P network.

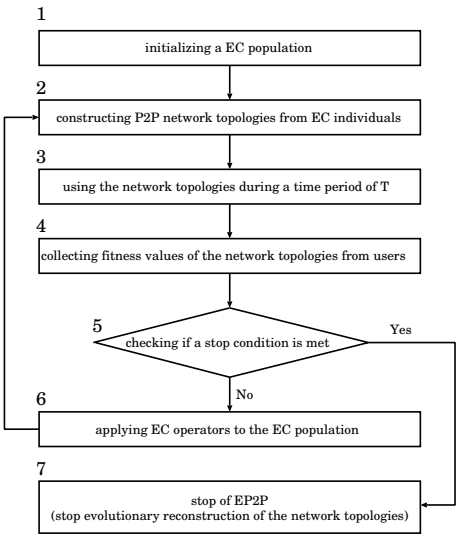


Fig. 1 A general flow chart of EP2P.

3.2 Information Retrieval

In L-EP2P, users issue queries, which propagate simultaneously over all co-existing topological structures of the P2P network by random walk. P2P network links among publishers (P2P nodes) are directed, so that the random walk randomly propagates the issued queries over a path tree that is uniformly determined. Each query is given an allowed number of hops (TTL), which is denoted as H_{max}. However, information found in some information source is delivered to a user who searched the network for that information through direct communication. In addition, it is assumed that all publishers can directly communicate with a super node.

Search queries indicate the contents of information held by information sources. For example, an information source that holds information on “movie” can respond to a search query of “movie”. In this case, as mentioned later herein, the information source (publisher) has shown what queries the information source can respond to and has also determined what queries the information source newly issues when the queries to which the information source can respond reach the information source.

When a search query reaches an information source that can respond to the search query within the allowed number of hops (H_{max}), the information source informs the user who issued the query of the location of the information source, and, at the same time, the information source issues a new search query. The newly issued query is propagated over the network by random walk. In this case, the newly issued query takes over the present hop counts and the allowed number of hops from the original search query. If the new query reaches an information source that can respond to the information source within the allowed number of hops, that information source behaves in the same manner as the first information source. In this way, new queries can be issued several times by different information sources. An example information search and retrieval process is shown in Figure 3. Also, a general flow chart of of information retrieval in L-EP2P is shown in Figure 4.

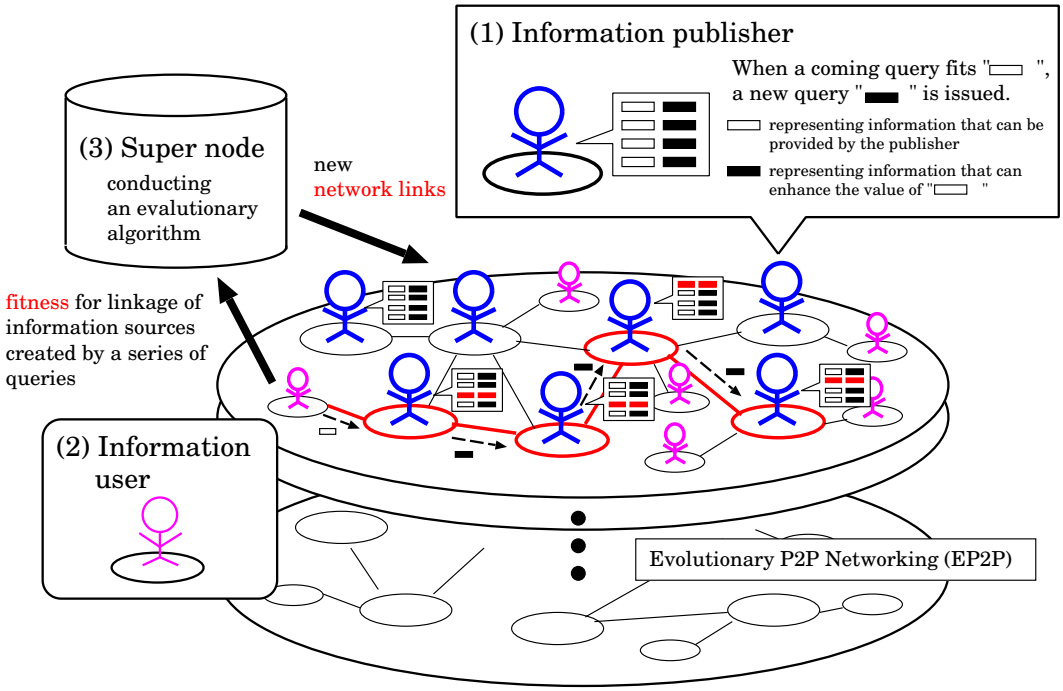


Fig. 2 Overview of the P2P information retrieval and sharing system.

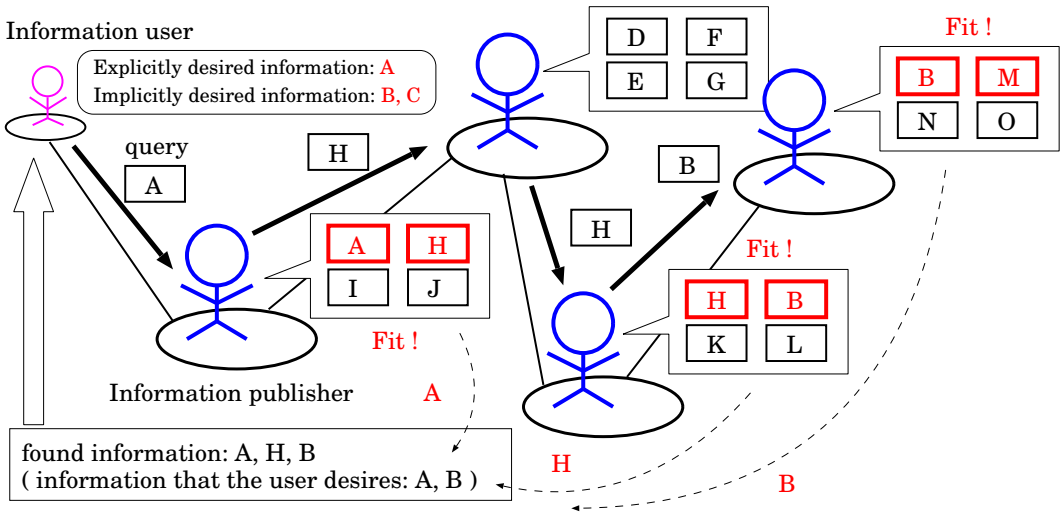


Fig. 3 Process of information retrieval in L-EP2P.

3.3 Proposal of Linkages between Information Sources

The proposal of linkage of information sources by a publisher occurs when the publisher makes new search queries that correspond to the search queries to which the publisher can respond. A pair of a respondable search query and its corresponding new search query is an instance of the linkage of information sources proposed by the publisher. In order to enhance its own information source, the publisher wants to have its own information source co-occur with an information source that corresponds to a new search query issued by the publisher. For example, suppose that the publisher holds information on “cooking recipes”. The publisher thinks that if an information source holding information on “cooking tools” co-occurs with its own information source as a result of a search for a user, then the value of its own information source becomes higher than the value when its own information source appears alone. In this case, the publisher prepares a pair of a respondable search query of “cooking recipes” and its corresponding new search query of “cooking tools”.

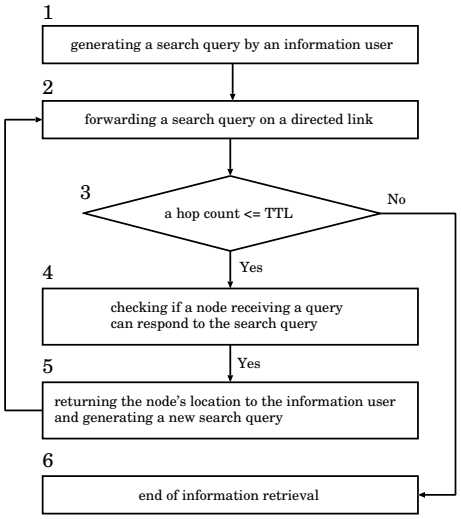


Fig. 4 A general flow chart of information retrieval in L-EP2P.

3.4 Evaluation of Linkages between Information Sources

Each user uses all of the present topological structures of the P2P network for a time period T and evaluates each topological structure, which is related to the linkage of information sources. The present topological structures are then reconstructed to new topological structures using the fitness provided by the users. This reconstruction of topological structures is conducted every time period T.

In one search for a specific information source by random walk, all of the topological structures are used. Whether an information source that is originally desired by a user issuing a search query is found and how many information sources that the user implicitly desires is found both depend on the shape of the topological structure. Therefore, a fitness is assigned to each topological structure, and this fitness is regarded as the number of information sources explicitly and implicitly desired by user that are found using the topological structure during time period T. In this case, each topological structure eventually has a certain fitness, and the topological structures with greater fitness can be considered to be better structures.

The above method for evaluating network topologies does not consider which information sources desired by users are included in each topological structure, but considers only how many information sources are included in it. Therefore, it is likely that topological structures that happened to include relatively more information sources desired by users become dominant in the population. That is to say, multiple topological structures are likely to include similar linkages among information sources. So, we will consider a following method for evaluating topological structures for the topological structures to include a variety of linkages among information sources. When a user conducts search, an identical search query is issued on N topological structures and then information sources desired by the user can be found in several topological structures. Then, the above-mentioned method uses only the number of desired information sources for fitness values. Meanwhile, in the new evaluation method considered here, when information source, S, desired by a user is found in N_s topological structures at the same time, each topological structure including the information source S is additively assigned a fitness value, f_s, expressed by Equation (1).

$$f_s = \frac{1}{r^{N_s - 1}}, \tag{1}$$

where r is a parameter. This method adds a small fitness value to a topological structure that provided a desired information source that was also provided by many other topological structures.

In L-EP2P, each topological structure is represented as the specific form presented in [7], and the specific form of the topological structure corresponds to an individual in the evolutionary algorithm that is used (see Figure 5). Each of L information source (P2P node) is assigned a serial number as its identifier, and the identifier corresponds to the index of a vector representing the individual. An element value of the individual (vector) represents an identifier of the information source to which a focus information source makes N_C directed links. A direction represented by a directed link indicates that a search query can be forwarded only in that direction.

Topological structures with better fitness basically include more linkages of information sources that were proposed by publishers and that users desired. Therefore, the linkages of information sources included in topological structures with better fitness are considered to be useful for both publishers and users.

3.5 Reconstruction of Network Topologies

Evolutionary operators are applied to the set of individuals mentioned above, which is referred to as a population, in order to generate a new set of individuals, which is referred to as the new population. The number of individuals held in the evolutionary algorithm, i.e., the population size, is N. Evolutionary operators generally include a selection operator, which is inspired by natural selection in Darwinism, a recombination or crossover operator, which models genetic recombination, and a mutation operator, which models gene mutation. The evolutionary operators used in the proposed EP2P are explained below.

3.5.1 Selection

The selection operator used herein is a tournament selection with a tournament size of K. The tournament selection ran-

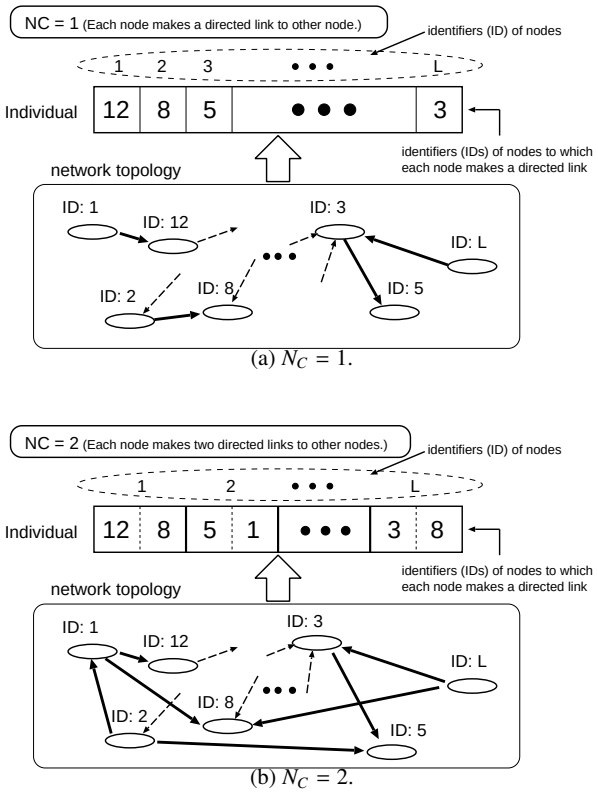


Fig. 5 Representation of a P2P network topology in the evolutionary algorithm (individual).

domly selects K individuals from the population and selects the individual with the best fitness among the K individuals. This selection procedure is repeated until N individuals have been selected.

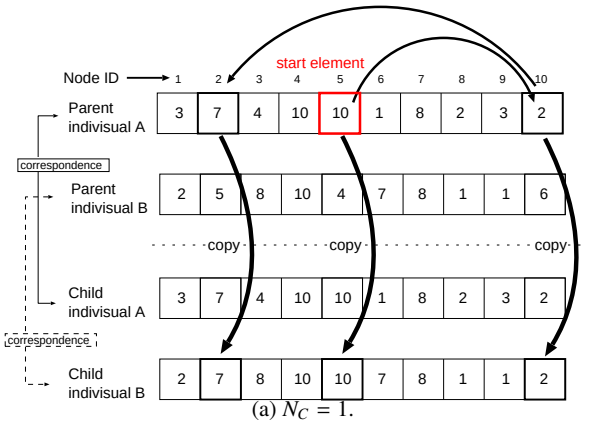
3.5.2 Crossover

The crossover operator used here is *node linkage crossover* (NLX) proposed in [7]. This operator is applied to an population as follows.

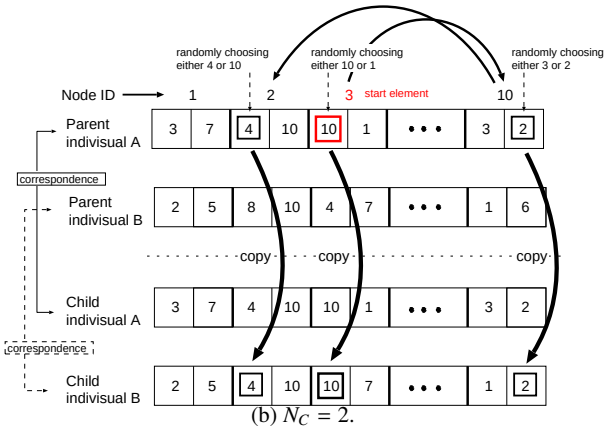
1. N individuals selected by the selection operator are divided into $N/2$ pairs of individuals. The selected individuals become parent individuals in this generation.
2. The crossover operator is applied to each pair of parent individuals with probability p_c. Child individuals generated from each pair of parent individuals are identical to the parent individuals before the crossover operator is applied. Each parent individual has a corresponding child individual.
3. For each pair of parent individuals to which the crossover operator is applied, one element is randomly selected from among the L elements of the individual. Recombination is conducted for the selected element with probability p_e.
4. For the element to which the recombination is to be applied, which child individual corresponding to one parent individual receives the element values of the other parent individual to be copied on itself is decided randomly.
5. After deciding which parent individual provides the element values for recombination, the node (element) linkage generated by directed links between nodes is copied to the target child individual.
   For example, suppose that $N_C = 1$, and the fifth element has been selected as the initial element of the linkage. Initially, NLX refers to the value of the fifth element of the parent individual as a copy source. If the reference value is 10, then NLX refers to the value of the tenth element. Furthermore, if the value of the tenth element is 2, then NLX refers to the value of the second element. By referencing the element values N_L times, NLX generates N_L element values and then copies them to the child individual corresponding to the other parent individual. In this example, N_L is 3, and the values of the reference elements are 5, 10, and 2, in that order. Nodes corresponding to the values of the reference element are linked by directed links. An example of this form of recombination is illustrated in Figure 6(a).
   Figure 6(b) shows an example of NLX with $N_C = 2$. In Figure 6(b), the third node has been selected as the initial node of the linkage. However, since each node makes two directed links, the third node has two elements that can be referred to by NLX, which, in this example, are 10 and 1. Then, NLX randomly chooses one of the two possible elements and refers to the value of the selected element, which is 10. Next, since the second node of the linkage, which is the tenth node, also has two elements, NLX randomly chooses one of the elements and refers to the value of the selected element, which is 2. In this way, the node linkage is formed. Generally, when $N_C \geqq 2$, NLX is performed in this manner.
6. Repeat Steps 3 through 5 $N_C \times L$ times.

3.5.3 Mutation

The mutation operator used herein is such that the value at each position (the gene) on the N individuals obtained after the node linkage crossover (NLX) is randomly changed to some other possible value with probability p_m, which is referred to as the mutation rate. This mutation operator is introduced mainly for bringing novel genes that did not appear in the initial population. In addition, if we set the mutation rate to be higher, the EP2P approaches to a random method.



(a) $N_C = 1$.



(b) $N_C = 2$.

Fig. 6 Example of node linkage crossover (NLX).

4 Simulations for Evaluation

4.1 Configurations

In the simulation model used herein, each publisher (node) corresponds to one information source. The corresponding information source for each publisher is determined from among K_I types of information sources in some way. In addition, the number of types of information sources with which each publisher wants to have its own information source co-occur is one, and this information source is determined from among K_I types of information sources in some way, excluding its own information source type.

In the simulation model, U users randomly link to information sources to make search queries. Every user implicitly searches the P2P network for any of the N_I types of information sources with one explicitly issued search query for a specific information source. When users explicitly issue a search query, they select one of the N_I types of information sources as the search query in some way. For example, for the case in which the number is $N_I = 1$, suppose that a user implicitly searches the network for information source of "B" with an issued search query of "A". In this case, the user does not explicitly express that the information source of "B" is a search target, but the user welcomes the finding of the information source of "B" as a result of searching for the information source of "A".

A unit of time is defined as a time period in which all users issue one search query and obtain the search results. A generation in the evolutionary algorithm used herein is considered to be T time units. In one generation, the present set of P2P network topologies, which are encoded into a population of the evolutionary algorithm, is used by all of the users for search, and at the end of the generation, the set of topologies are reconstructed by the evolutionary operators for the next generation. One simulation run lasts 50 generations ($50\times T$ time units).

L-EP2P is compared to a P2P network for information sharing that includes the same number of network topologies as L-EP2P but does not reconstruct the topologies. Through this comparison, we examine the impact of the reconstruction of the network topologies, i.e., the reconstruction of the linkages of information sources by EP2P.

The parameter values for the simulation model used herein as well as the parameter values for L-EP2P used in the simulations are shown in Table 1.

4.2 Evaluation Scenarios

First, the common things to all evaluation scenarios considered here are as follows. The initial network topologies are randomly generated. Node departure and participation does not occur during a run of simulation. At every unit time, every user randomly selects an information source and then issues a search query only once from the selected information source.

A simulation scenario specifies how to decide a type of information source for each information source and a type of information source that the information source proposes as a co-occuring information source with itself among K_I types of information sources and also how to decide a type of information source that a user explicitly searches and N_I types of information sources that the user implicitly desires among K_I types of information sources. Specifically, we prepare the following four evaluation scenarios.

1. This evaluation scenario randomly determines a type of information source for each information source and a type of information source that the information source proposes as a co-occuring information source with itself among K_I types of information sources (referred to as **information source's type** hereinafter) and a type of information source that a user explicitly searches and N_I

Table 1 Parameter values.

parameters	description	values
L	The number of publishers	500
U	The number of information users	2,000
N_C	The number of directed links that an information publisher makes	2
N	The number of topological structures of a P2P network	50
T	A time period during which the present topological structures are used	20
K_I	The number of sorts of information sources as search objects	30
N_I	The number of sorts of information sources that an information user implicitly desires in one search	from 1 to 4
H_{max}	The number of hops allowed for one search	10
K	The tournament size for the tournament selection	2
p_c	The crossover rate	1.0
p_e	The exchange rate between genes in the node linkage crossover (NLX)	0.005
N_L	The length of linkage of exchanged genes in the NLX	5
p_m	The mutation rate	0.1
r	The parameter for determining the fitness value among topologies including the same desired information sources	4

types of information sources that the user implicitly desires among K_I types of information sources (referred to as **information user's type** hereinafter). The information source's type is randomly determined at the beginning and fixed after that. Meanwhile, the information user's type is randomly determined at every search.

2. The information source's type is randomly determined and the information user's type is determined following Zipf's Law [17]. Zipf's law is represented by Equation (2).

$$f = kx^{-\alpha}, \tag{2}$$

where f is the selection ratio for each type of information source, x is the rank of popularity of each type of information source and also equivalent to the serial number that is uniquely assigned to each type of information source, and α stands for the degree of imbalance of popularity among all information sources. According to Zipf's law, a few information sources with high popularity have most accesses. In this section, we will set the parameter $\alpha = 1.0$ in Equation (2). The query distribution following Zipf's law with $\alpha = 1.0$ is shown in Figure 7.

3. The information source's type is determined following Zipf's Law and the information user's type is randomly determined.
4. Both information source's type and information user's type are determined following Zipf's Law.

For all the four evaluation scenarios mentioned above, two types of methods for calculating fitness values of network topologies, which were described in Section 3.4, are used in the evolutionary algorithm. We will hereinafter refer to the evaluation method that counts the accumulative number of desired information sources found in each network topology as a fitness value of the network topology as **the simple evaluation way**, and also refer to the evaluation method that additively assigns a smaller fitness value to multiple network topologies in which the same type of

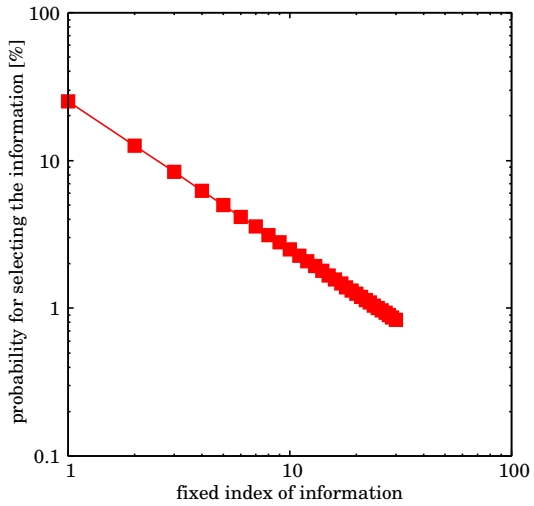


Fig. 7 The query distribution following Zipf's law with $\alpha = 1.0$.

desired information source was found as **the complicated evaluation way**.

4.3 Observation Items

In L-EP2P, a user making a search query concerns how many information sources that the user explicitly and implicitly desires were eventually obtained, and an information source (information publisher) concerns how much linkages among information sources that it proposed were evaluated by users. Both concerns can be examined by observing how many desired information sources were found in a trial of search. Therefore, we set the first observation item to be change in the average number of obtained desired information sources in a trial of search over time period of T. We will refer to this first observation item as **the average number of obtained information sources** hereinafter. Since all network topologies are used for a trial of search, a user can obtain a desired information source if the desired information source

is found in either of all network topologies. If the identical type of desired information source is found in several network topologies, then the number of obtained information sources is counted as one.

The second observation item is change in the average number of new search queries issued by information sources in a trial of search over time period of T. We will refer to this second observation item as **the average number of new search queries**. We can roughly grasp how many proposals of information sources were accepted by users by comparing the first and the second observation items. If the identical search queries are newly issued in several network topologies, then each issued query is counted in the number of new search queries.

4.4 Results and Discussion

The first observation item, that is the average number of obtained information sources for the four evaluation scenarios is shown in Figure 8. The second observation item, that is the average number of new search queries for the four evaluation scenarios is shown in Figure 9. For every evaluation scenario, four types of N_I values, $1, 2, 3, 4$, were used. In addition, three types of topology reconstruction methods were used, which are the evolutionary topology reconstruction method using the simple evaluation way, whose results are labeled "simple evaluation", the evolutionary topology reconstruction method using the complicated evaluation way, whose results are labeled "complicated evaluation", and the no topology reconstruction, whose results are labeled "no evolution". All the results were the average over ten independent runs. Here, the initial observed values for two types of the evolutionary reconstruction methods and those for no reconstruction are different, as shown in Figure 8. However, that is because seeds for the used random number generator are different among them and the initial differences do not represent the difference in their performances. How the initial values change is a matter.

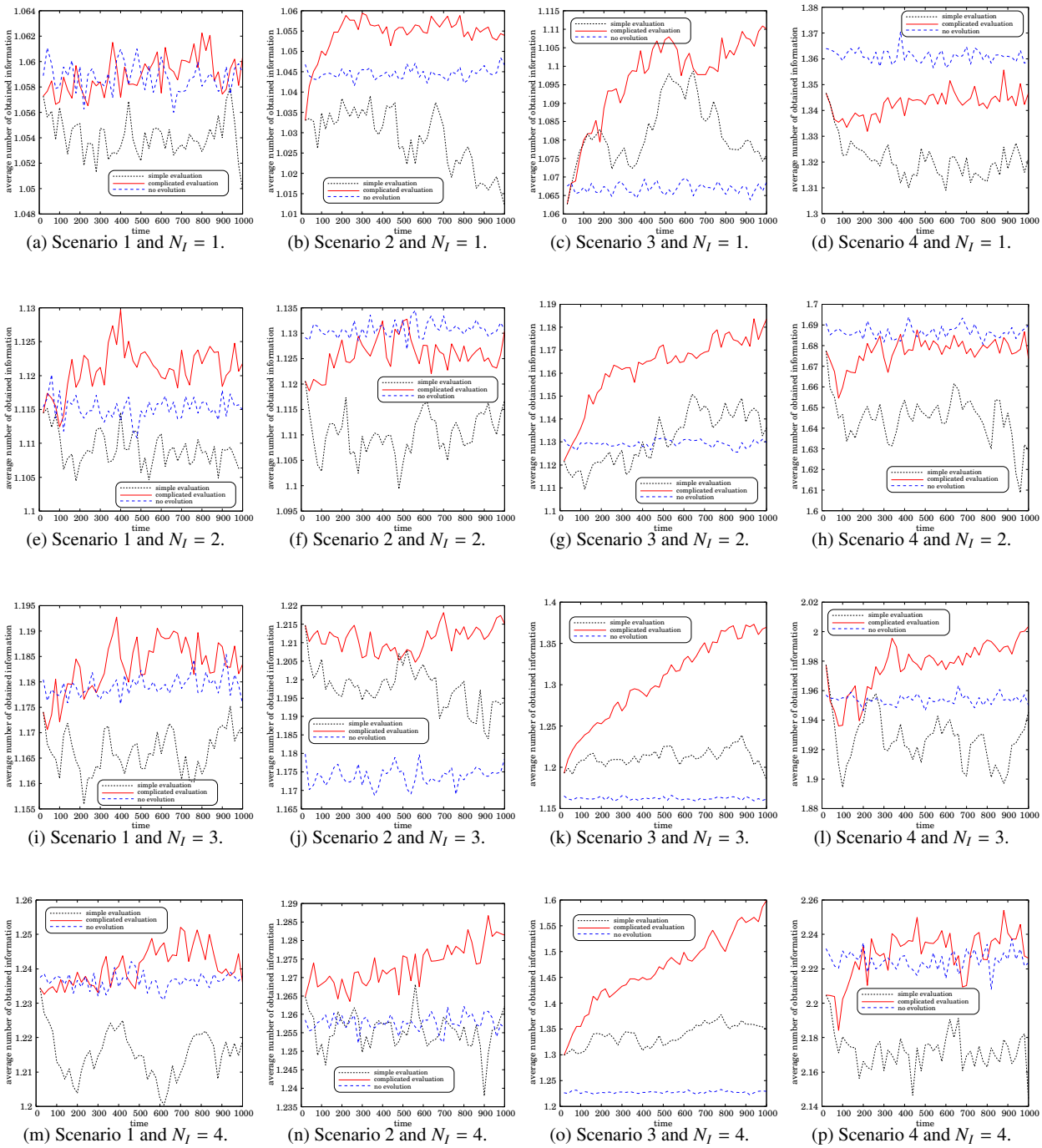
We can observe from Figure 8 that each evaluation scenario shows its own tendency in change of the average number of obtained information sources no matter what number of types of information sources that a user implicitly desires in one search, N_I, is used. In the evaluation scenario 1, the simple evaluation way decreased the average number of obtained information sources with the time a little, the complicated evaluation way increased that with the time a little, and no reconstruction had almost the same number of obtained information sources during the time period. The evaluation scenario 2 has a similar tendency in change of the average number of obtained information sources with the evaluation scenario 1. In addition, the evaluation scenarios 1 and 2 have similarity in the average number of obtained information sources itself. In the evaluation scenario 3, both simple and complicated evaluation ways increased the average number of obtained information sources with the time, and the average number of obtained information sources for the complicated evaluation way is larger than that for the simple evaluation way. The evaluation scenario 4 also has the similar tendency with the evaluation scenarios 1 and 2. However, the average number of obtained information sources for the evaluation scenarios 3 and 4 is larger than that for the evaluation scenarios 1 and 2. Thus, the evolutionary topology reconstruction method with the complicated evaluation way can be regarded as the best one because it could increase the average number of obtained information sources for any evaluation scenario with any value of N_I. Meanwhile, the evolutionary topology reconstruction method with the simple evaluation way can be said to be less effective than other methods because it was often inferior to the no topology reconstruction.

Next, from Figure 9, we can observe that there is a common tendency in change of the average number of new queries among all of the evaluation scenarios. The common tendency is that the simple evaluation way yielded the largest increase of the average number of new queries, the complicated evaluation way yielded the second largest increase of the average number, and no evolution had a roughly constant average number over the time period.

As mentioned above, L-EP2P with the simple evaluation way is worse even than the no topology reconstruction except in the evaluation scenario 3, but yielded the largest average number of new queries for any evaluation scenario. First, we will discuss reasons for that below.

L-EP2P with the simple evaluation way, as mentioned in Section 3.4, assigns high fitness values to network topologies in which more desired information sources were provided for users and does not consider what information source was provided in each network topology. Therefore, network topologies similar to one that was found having a relatively high fitness value in the evolutionary process occupy a large part of the population. Such occupancy of similar network topologies in the population indicates that network topologies that give easy access to particular information sources overlap in the population. In other words, particular search queries for the particular information sources are issued in multiple network topologies in parallel in a trial of search. Therefore, the average number of new queries keeps increasing in L-EP2P with the simple evaluation way, as shown in Figure 9. At the same time, multiple network topologies become specialized to easily access to only particular information sources. That is to say, diversity of network topologies gets lost and then the average number of obtained information sources keeps decreasing, as shown in Figure 8.

L-EP2P with the complicated evaluation way, as mentioned in Section 3.4, assigns a high fitness value to a network topology that provides desired information sources hard



(a) Scenario 1 and $N_I = 1$. (b) Scenario 2 and $N_I = 1$. (c) Scenario 3 and $N_I = 1$. (d) Scenario 4 and $N_I = 1$.

(e) Scenario 1 and $N_I = 2$. (f) Scenario 2 and $N_I = 2$. (g) Scenario 3 and $N_I = 2$. (h) Scenario 4 and $N_I = 2$.

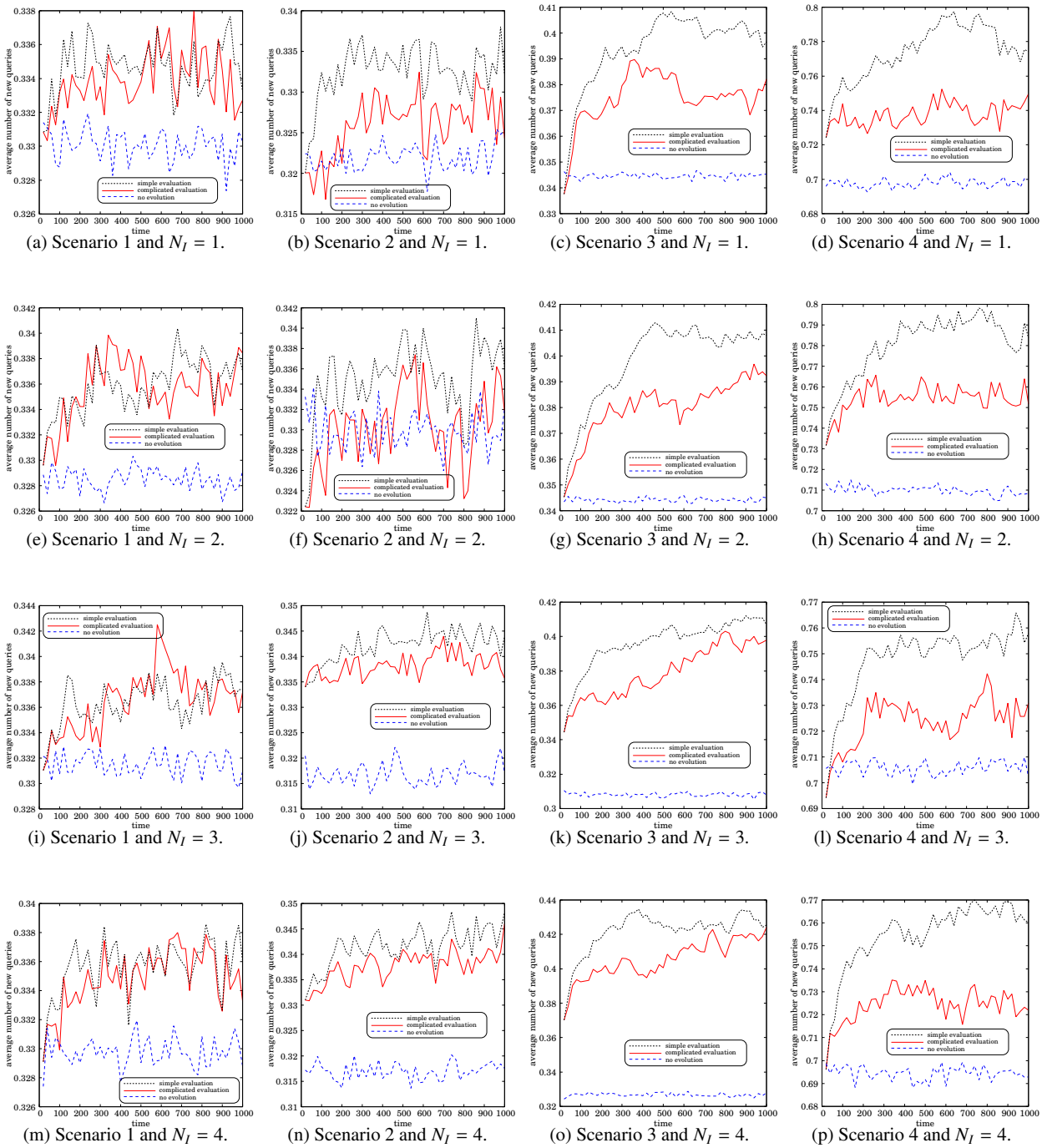
(i) Scenario 1 and $N_I = 3$. (j) Scenario 2 and $N_I = 3$. (k) Scenario 3 and $N_I = 3$. (l) Scenario 4 and $N_I = 3$.

(m) Scenario 1 and $N_I = 4$. (n) Scenario 2 and $N_I = 4$. (o) Scenario 3 and $N_I = 4$. (p) Scenario 4 and $N_I = 4$.

Fig. 8 Change in the average number of obtained desired information sources in a trial of search over time period of T.

to obtain in other network topologies for users and a low fitness value to a network topology that provides desired information sources easy to obtain in other network topologies for users. Therefore, unlike L-EP2P with the simple evaluation way, it can maintain diversity of network topologies. As a result, as shown in Figure 9, it has smaller increase of the average number of new search queries but larger increase of the average number of obtained information sources than L-EP2P with the simple evaluation way.

Next, we will focus on each evaluation scenario and discuss reasons why each evaluation scenario showed its own tendency in change of the average number of new queries as well as the average number of obtained information sources below.



(a) Scenario 1 and $N_I = 1$. (b) Scenario 2 and $N_I = 1$. (c) Scenario 3 and $N_I = 1$. (d) Scenario 4 and $N_I = 1$.

(e) Scenario 1 and $N_I = 2$. (f) Scenario 2 and $N_I = 2$. (g) Scenario 3 and $N_I = 2$. (h) Scenario 4 and $N_I = 2$.

(i) Scenario 1 and $N_I = 3$. (j) Scenario 2 and $N_I = 3$. (k) Scenario 3 and $N_I = 3$. (l) Scenario 4 and $N_I = 3$.

(m) Scenario 1 and $N_I = 4$. (n) Scenario 2 and $N_I = 4$. (o) Scenario 3 and $N_I = 4$. (p) Scenario 4 and $N_I = 4$.

Fig. 9 Change in the average number of new search queries issued by information sources in a trial of search over time period of T.

The evaluation scenario 1 randomly determines the information source's type as well as the information user's type. In addition, an information source from which a user first issues a search query is also randomly determined. Therefore, network topologies should not be specialized for searching specific information sources from specific locations. However, the information source's type is fixed after the random initialization at the beginning, so that there is possibility that the fixed information source's type happens to include bias that brings advantages or disadvantages in searching specific information sources. In that case, it is expected that evolutionary change in the initially fixed network topologies sometimes improves the search performance. The simulation results shows that L-EP2P with the simple evalua-

tion way is the worst. That is, as mentioned above, because it is not capable of maintaing diversity of network topologies. L-EP2P with the complicated evaluation way can contrastively maintain the diversity. Furthermore, unlike the no topology reconstruction, L-EP2P with the complicated evaluation way has opportunities to evolutionarily change network topologies with some bias, and therefore, outperforms the no topology reconstruction slightly.

The evaluation scenario 2 determines the information source's type randomly and also determines the information user's type following Zipf's Law, which means that information sources desired by users are biased. Since searching for particular information sources begins from a randomly determined information source, information sources that can respond to search queries for the particular information sources should be uniformly placed over the network. Then, since a type of information source that an information source has is randomly determined, there are not so many information sources that can respond to the search queries for the particular information sources. Therefore, random construction of network topologies is basically the best method. However, the information source's type is first randomly determined and then fixed, so that it is, as in the evaluation scenario 1, expected that evolutionary change for relaxing some bias included in the initial topologies contributes to improving the performance of L-EP2P. Therefore, the tendency of the results for the evaluation scenario 2 is similar to that for the evaluation scenario 1.

The tendency of the results for the evaluation scenario 4 is also similar to that for the evaluation scenarios 1 and 2. So, the randomly constructed network topologies are basically the best. However, the average number of obtained information sources for the evaluation scenario 4 is larger than that for the evaluation scenarios 1 and 2. The reason for that would be that types of information sources desired by users and types of information provided by information sources are both biased but matched.

Finally, only the results of the evaluation scenario 3 show different tendency from the other scenarios'. In the evaluation scenario 3, not only L-EP2P with complicated evaluation way but L-EP2P with the simple evaluation way as well show the better performance than the no reconstruction. That indicates that the evolutionary topology reconstruction is more effective than no topology reconstruction. In addition, L-EP2P with complicated evaluation way is better than L-EP2P with the simple evaluation way. That indicates that the evolutionary topology reconstruction maintaining diversity of the population is more effective than that loosing the diversity. The evaluation scenario 3 determines the information source's type following Zipf's Law and determines the information user's type randomly. Then, since types of information sources that information sources have are biased, searching for information sources that many information sources have is reliable. Meanwhile, randomly determined network topologies cannot reliably respond to search queries issued from all around the network for information sources that the less information sources have. In order for L-EP2P to satisfy users' requests in such a situation, it is necessary to make the less information sources be easily accessed. Therefore, it can be thought that the evolutionary topology reconstruction from random network topologies to such network topologies is effective. The performance difference between L-EP2P with the simple and the complicated evaluation ways would come from the difference in ability in maintaining diversity of the population.

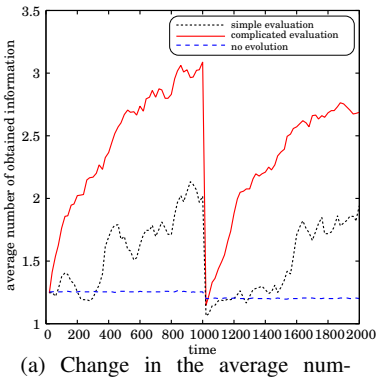
4.5 Adaptability to Dynamic Change in Environments

In the evaluation of L-EP2P described in the previous section, we did not consider change in environments such as query distribution and node participation and departure (node churn). Here we examine adaptability of L-P2P to large changes in query distribution and node churn. We use only $N_I = 4$ here.

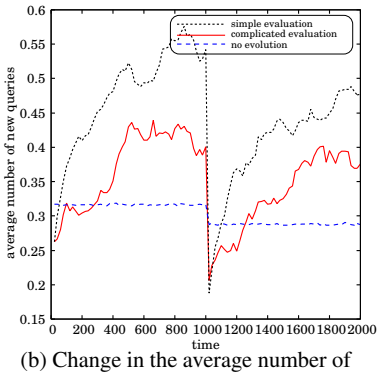
First, we make a model for a large change in query distribution as follows. The information source's type is randomly determined. Meanwhile, the information user's type is the same for all users but changed to completely different one at the middle of a simulation run, where the total time period of the simulation run is 2000. Since all users desire the same types of information sources in this model, it is likely that network topologies converge to particular ones before the large change of query distribution. Then, users' demands suddenly and completely change, so that quick adaptation of the network topologies to the change is required.

Next, we make a model for node churn as follows. Each of the L nodes decides whether it will join the network at each time unit according to its given probability. This probability is hereinafter referred to as the participation probability. The participation probability of each node is determined as a uniform random real number in $[P_L, 1]$. Each node joins the network with its participation probability. If a node does not join the network, then the node is in the state whereby the node leaves the network. After all of the nodes make a decision with regard to participation, each of the nodes conducts one search. A time unit is regarded as the period of time required for all of the nodes to make this decision and complete one search. Here we use three values, $0.95, 0.8, 0.5$, as P_L. The smaller the value of P_L, the higher the node churn.

Figure 10 shows results on adaptability of L-EP2P to change in query distribution. In this figure, as the result graphs presented in the previous section, two types of observed values are drawn, which are change in the average number of obtained desired information sources and change in the average number of new search queries. Also, Figure 11 shows



(a) Change in the average number of obtained desired information sources.



(b) Change in the average number of new search queries.

Fig. 10 Adaptability of L-EP2P to change in query distribution.

results on adaptability of L-EP2P to node churn. All the results were the average over ten independent runs.

We can observe from Figure 10 that the evolutionary topology reconstruction method could adapt the network topologies to the change in query distribution occurred at time 1000. Meanwhile, the no topology reconstruction method has almost no change in the number of obtained desired information sources and the number of new search queries. In addition, Figure 10 shows that the evolutionary topology reconstruction method with the complicated way is better than that with the simple way in terms of adaptation ability. In this simulation scenario, L-EP2P is required to make it easier to find the identical information sources that all of users desire from any user (any network location). Therefore, as in the simulation scenarios presented in the previous section, maintaining diversity of network topologies is needed to achieve better adaptation, and the complicated way, which is superior to the simple way with respect to ability in maintaining diversity of network topologies, has higher adaptation ability than the simple way.

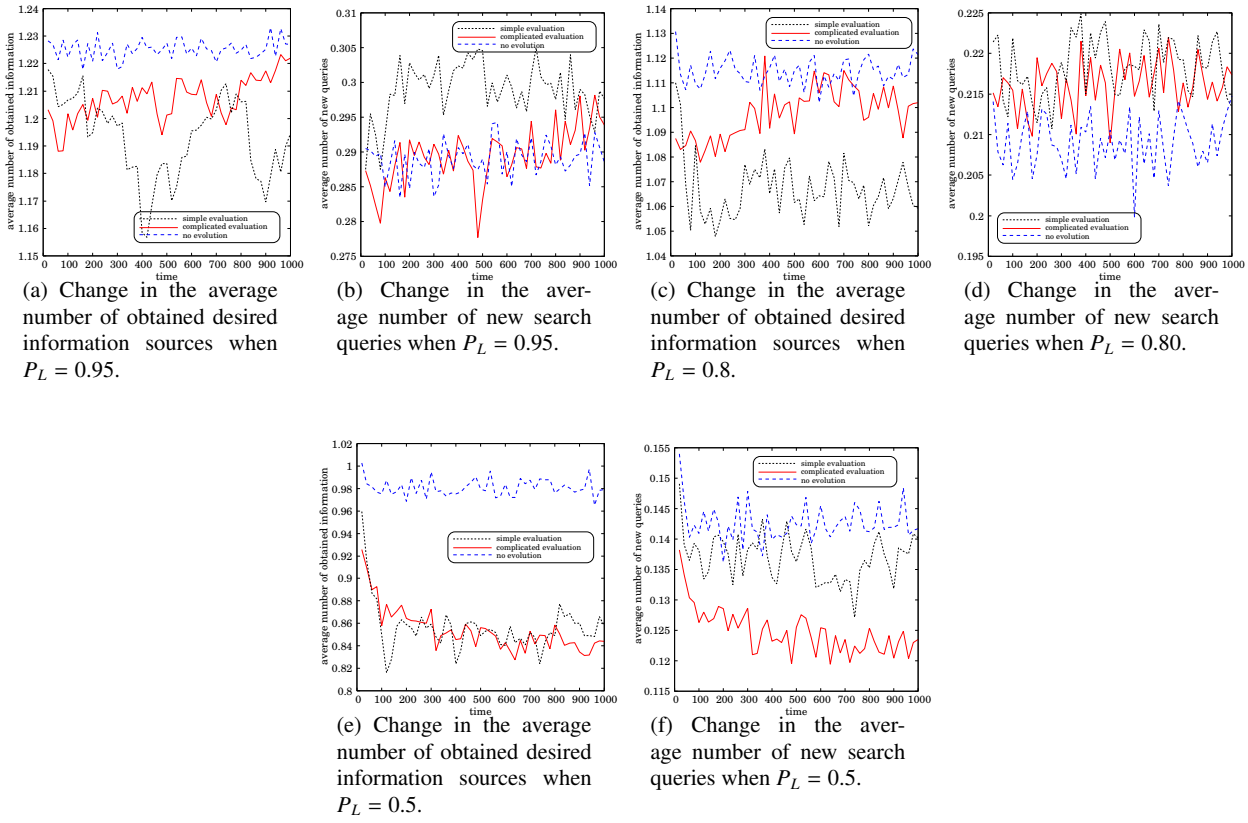
The change in query distribution considered here brings a drastic change in the fitness function used by EP2P. Furthermore, since P2P nodes from which users issue a search query are randomly determined though the desired information sources are the same for all the users, the fitness function before and after the change in query distribution is not completely static. However, the fitness function before and after the change in query distribution is considered to be almost static, and in such a situation, we can expect that evolutionary adaptation works well, that is to say, we can expect that individuals fitting to the present environments also fit to future environments.

Meanwhile, when nodes participating in the network change every time, the fitness function also changes every time. The degree of the change in the fitness function depends on how much participating nodes change every time. In the simulation scenario considered here, the change in the fitness function becomes bigger as the value of P_L becomes smaller. In this case, evolutionary adaptation would be basically hard to occur and the simulation results actually supports this expectation. Figure 11 indicates that when the the change in the number of participating nodes every time becomes bigger (the value of P_L becomes smaller), the adaptation ability of the evolutionary topology reconstruction method becomes lower. When P_L is 0.95 or 0.8, the evolutionary topology reconstruction method with the complicated way shows adaptation ability somewhat, but when P_L is 0.5, the two types of the evolutionary topology reconstruction method is worse than the no topology reconstruction.

As shown in Figure 11, evolutionary adaptation is hard to occur when the change in participating nodes (participating information sources) is drastic and frequent. However, if we consider L-EP2P to be a network that is composed of publishers who stably publish their own information sources to users, we do not need to consider a frequent and drastic change in participating nodes, that is, high node churn. Meanwhile, we could think it reasonable that users freely leave and join the network. If users who have so different demands appear in the network every time, the fitness function can frequently and drastically change, and therefore, evolutionary adaptation of the network topologies would be hard to occur. We need to consider a more realistic model for the users and investigate the proposed system using the model.

5 Conclusions

The present paper proposed an information retrieval and sharing system, referred to as L-EP2P, that creates linkages of information sources that are useful for both information publishers and users in a P2P network. Simulation results revealed that L-EP2P could create more useful linkages of information sources for both information publishers and users,



(a) Change in the average number of obtained desired information sources when $P_L = 0.95$.

(b) Change in the average number of new search queries when $P_L = 0.95$.

(c) Change in the average number of obtained desired information sources when $P_L = 0.8$.

(d) Change in the average number of new search queries when $P_L = 0.80$.

(e) Change in the average number of obtained desired information sources when $P_L = 0.5$.

(f) Change in the average number of new search queries when $P_L = 0.5$.

Fig. 11 Adaptability of L-EP2P to node participation and departure (node churn).

as compared to a P2P network without topology reconstruction. More concretely, a situation in which the evolutionary topology reconstruction works effectively is that users issue search queries for less types of information sources from various locations. In addition, maintaining diversity of the population is needed to achieve better performance of L-EP2P when the evolutionary topology reconstruction is used. Moreover, even in a situation that random topology construction is basically the best effective, the evolutionary topology reconstruction can slightly improve the performance by alleviating the bias that happened to be included in randomly constructed topologies. Furthermore, the evolutionary topology reconstruction is shown to be effective when the node churn (change in participating nodes) is low, that is, when the fitness function is somewhat static.

In the future, we will evaluate L-EP2P in more realistic environments through simulations and will compare L-EP2P to other approaches to creating linkages of information sources.

References

1. Delicious. http://delicious.com/
2. Flickr. http://www.flickr.com/
3. A. Mathes, "Folksonomies - Cooperative Classification and Communication Through Shared Metadata", http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html
4. E. K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim, "A Survey and Comparison of Peer-to-Peer Overlay Network Schemes", IEEE Communications Surveys & Tutorials, Vol. 7, No. 2, pp.72–93, 2005.
5. M. Ripeanu, A. Iamnitchi, and I. Foster, "Mapping the gnutella network," *IEEE Internet Computing*, vol. 6, no. 1, pp. 50–57, January/February 2002.
6. D. Stutzbach, R. Rejaie, and S. Sen, "Characterizing Unstructured Overlay Topologies in Modern P2P File-sharing Systems," *IEEE/ACM Transactions on Networking (TON)*, vol. 16, no. 2, pp. 267–280, 2008.
7. K. Ohnishi and Y. Oie, "Evolutionary P2P Networking that Fuses Evolutionary Computation and P2P Networking Together," IEICE Transactions on Communications, Vol. E93-B, No. 2, pp.317–328, 2010.
8. T. Bäck, *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford University Press, 1996.
9. S. G. M. Koo, C. S. G. Lee, and K. Kannan, "A Genetic-algorithm-based Neighbor-selection Strategy for Hybrid Peer-to-peer Networks," In Proceedings of the International Conference On Computer Communications and Networks (ICCCN 2004), pp. 469–474, 2004.
10. M. Srivatsa, B. Gedik, and L. Liu, "Large Scaling Unstructured Peer-to-peer Networks with Heterogeneity-aware Topology and

Routing," IEEE Transactions on Parallel and Distributed Systems, Vol.17, No.11, pp.1277–1293, November 2006.
11. E. Pournaras, G. Exarchakos, and N. Antonopoulos, "Load-driven Neighbourhood Reconfiguration of Gnutella Overlay," Computer Communications, Vol.31, No.13, pp.3030–3039, August 2008.
12. P. Merz and S. Wolf, "Evolutionary Local Search for Designing Peer-to-peer Overlay Topologies based on Minimum Routing Cost Spanning Trees," In Proceedings of the 9th International Conference on Parallel Problem Solving from Nature (PPSN IX), pages 272–281, 2006.
13. M. Munetomo, Y. Takai, and Y. Sato, "An Adaptive Network Routing Algorithm Employing Path Genetic Operators," Proceedings of the Seventh International Conference on Genetic Algorithms, pp.643–649, 1997.
14. P. Imai and C. Tschudin, "Practical Online Network Stack Evolution," SASO 2010 Workshop on Self-Adaptive Networking, September 2010.
15. D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry", Communications of the ACM, Vol. 35, No. 12, pp. 61–70, 1992.
16. D. E. Goldberg,"Genetic Algorithms in Search, Optimization, and Machine Learning," Addison-Wesley, 1989.
17. LA. Adamic and BA. Huberman, "The Nature of Markets in the World Wide Web", Quarterly Journal of Electronic Commerce, Vol.1, No.1, pp.5–12, 2000.